UNIVERSITY OF
Southampton
School of Electronics
and Computer Science

Research Review

Research Methods in Computing

A Survey of Ontology Construction and Information Extraction from Wikipedia

Author:

Chaohai Ding
Email:cd8e10@ecs.soton.ac.uk

Supervisor:

Dr. Srinandan
Dasmahapatra

December 10, 2010

Contents

Abstract

With the development of the Semantic Web in recent years, the ontology has become a significant factor for information extraction and inference, which contributes to the semantic search. Wikipedia known as the largest corpus for information extraction plays an essential role in machine learning, data mining and semantic search. Also a well-formed ontology plays a crucial role in the semantic search. This paper gives a brief introduction about the ontology construction and compares some existing systems for refining and reconstructing the Wikipedia.

1 Introduction

An ontology was defined as "an explicit, machine-readable specification of a shared conceptualization" in [1]. Wikipedia known as the largest corpus for information extraction plays an essential role in machine learning, data mining and knowledge-based system [2]. But the current Wikipedia ontology such as the Infobox ontology was mainly constructed manually by volunteers' collaboration and some articles of Wikipedia are redundant and ambiguous [3]. So in this report, section one presents a historical overview about the ontology refining and extraction, while section two summarizes the three key publications in this area following with evaluation of these publications in section three.

2 Historical Overview

1993, Gruber, T.R. et al. [4] summarized an approach to the ontology specification. Then following with Studer et al.[5] described some roles of ontology in different areas and presented a brief introduction about constructing ontology such as constructing ontology form other ontologies and internal integration of an ontology, which contributes to information services and knowledge management in 1998, Gomez-Perez and Benjamins [6] presented an approach to reuse ontologies in 1999. [7] described a methodology using frame-based system to develop the ontology in 2001, while in 2002 Chandrasekaran et al. [8] presented the significant role of an ontology in knowledge technology and illustrated a way to describe the world with ontologies. In 2005, Maedche and Staab [9] presented an ontology-learning framework which is using semiautomatic ontology-construction tools to refine the ontology. During this period, some ontologies edited manually come forth continuously, such as a lexical English database called WordNet [10], SUMO [11] and Cyc [12].With the drastic development of Wikipedia and Social Networks, experts in the area of Artificial Intelligent and Knowledge Based System commence to use such huge information as ontologies to automatically reconstruct new ontologies which would contribute to the development of the Semantic Web, such as Kylin [3], YOGO [13]and DBpedia [14].

3 Summaries of Key Publications

Wikipedia is not just a platform edited by volunteers collaboratively [15], but also a large ontology contributing to the development of semantic search [16, 17], machine translation [18]and word sense disambiguation [19]. Summaries of three main publications would be presented in this section. These three respectively introduce an extension to integrate Wikipedia, a new ontology generated from Wikipedia and an approach to refine Wikipedia then extract information from it.

3.1 Semantic Wikipedia

[20] describes an extension to integrate Wikipedia through reviewing the related approaches and presents a basic structure of the current implementation. The project related to this paper mainly invites volunteers who are authors of Wikipedia to refine the collaboratively edited information from Wikipedia into standardization and machine readable using RDF [21],

OWL [22] and RDFS [23]. However, the Sematic MediaWiki Project is a comparable initiative project to structure Wikipedia into semantic network. This paper gives a detail about attributes and usage of typed links in Wikipedia. Related approaches in semantically enhanced Wikipedia are presented in a wide range, which would give an overview about the structure of the Wikipedia Ontology. Also, this implementation would benefit for the semantic technology developers who use a huge machine accessible data.

3.2 YAGO: A Large Ontology from Wikipedia and WordNet

[24] evaluates some existing approaches to construct an ontology and presents a high quality approach "type checking" to construct an ontology from Wikipedia combining with WordNet automatically. With the high precision of 95% in the unification between Wikipedia and WordNet, YAGO [13] shows an advanced performance in ontology construction. A new approach of information extraction is presented, which includes entities extracted from Wikipedia and the implement of quality control. The techniques known as reductive and inductive type checking can be applied in other corpora. YAGO as a large reconstructed ontology provides the semantic source for the Semantic Web as well as contributes to other information extraction projects.

3.3 Automatically Refining the Wikipedia Infobox Ontology

[3] presents an approach to refine the Wikipedia Infobox Ontology using the machine learning system known as Kylin Ontology Generator (KOG) [25] which constructs an abundant ontology by combining with the WordNet to generate the well-formed ontology automatically. It mainly addresses the problem of refining ontology and identifies the aspects of Wikipedia data source through classification. The experiment result shows an effective improvement in advanced query such as a SQL-like question and contributes to a rich ontology which would enhance the Wikipedia Infobox Ontology. According to this kind refining ontology, people could make an advanced query and then the system would response a parallel set of answers. Moreover, using the refined Wikipedia Infobox Ontology as the training data for machine learning system would benefit for next-generation question answering system and semantic web search [26].

4 Evaluation

Figure 1: Ontology in Semantic MediaWiki [27]

[20] improves the Wikipedia's capabilities of semantic search with the standardization of Semantic Web Technology. In aspect of man-made ontology, it initiatively invites volunteers to edit Wikipedia with semantic tagging collaboratively. Semantic MediaWiki, the project related to this paper mainly addresses the problem that the context of Wikipedia is not machine-readable in the early year and establishes an extensive knowledge base which benefits for semantic web. Compared with WordNet, Semantic MediaWiki provides a wide range of objects, but the manually edited ontology leads to the heavy cost of workforce which should be supported by communities. Moreover, the quality of target ontology also should be controlled, such as up to date information. Figure 1 indicates the information lag in the population of London.

Compared with Semantic MediaWiki, YAGO and Kylin automatically refine and construct new ontology from existing Wikipedia Infobox Ontology. YAGO as a part of Linking Open Data project [28] performs a high precision in constructed ontology. But YAGO must be based on YAGO model which expresses the connection between entities and facts. It uses heuristics to extract information from the knowledge base which combines Wikipedia and WordNet. Moreover, the quality controlling applied into this system provides an advanced query and semantic search. But YAGO constructs the ontology in relatively narrow area compared with other ontology such as DBpedia. Kylin is similar with DBpedia, which refines the existing ontology source combined Wikipeida and WordNet through powerful machine

learning system automatically. This system addresses the problems that Wikipedia Infobox Ontology is not well-defined and redundant. Automatically refined ontology and high accuracy of information extraction could contribute to semantic search. Figure 2 indicates the precision of information extraction in different ontologies.

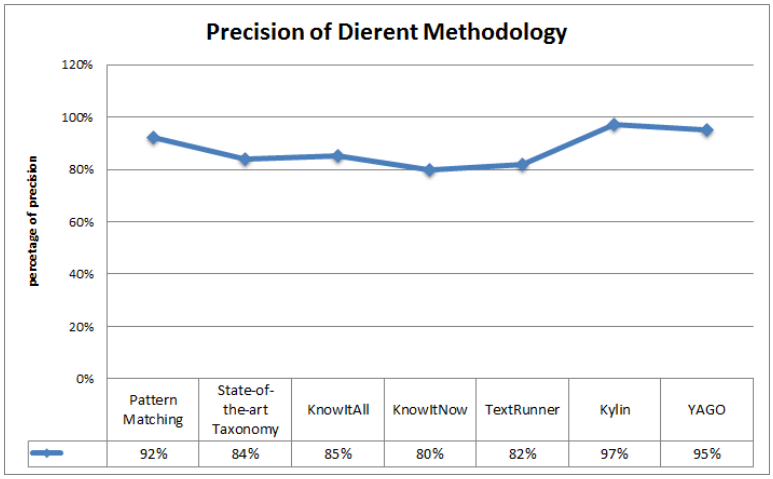

Figure 2: Precision of Different Methodology [24]

These three projects all construct the ontology from Wikepedia, Semantic MediaWiki provides a huge semantic network using manual collaboration. Kylin produces a well-defined ontology from Wikipedia Infobox Ontology through machine learning system automatically, while YAGAO provides a large ontology constructed from Wikipedia and performs a high precision in information extraction.They have different methodology, but the same goal is providing an approach to extend or reconstruct the ontology from existing ontology such as Wikipedia.

5 Bibliography

[1] N. Guarino and I. (Roma) Consiglio nazionale delle ricerc, "Formal ontology in information systems," 1998.

[2] K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," *Web Information Systems Engineering–WISE 2007*, pp. 322–334, 2007.

[3] F. Wu and D. Weld, "Automatically refining the Wikipedia infobox ontology," in *Proceeding of the 17th international conference on World Wide Web*. ACM, 2008, pp. 635–644.

[4] T. Gruber *et al.*, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, pp. 199–199, 1993.

[5] R. Studer, V. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," *Data & knowledge engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.

[6] A. Gomez-Perez and V. Benjamins, "Applications of ontologies and problem-solving methods," *AI Magazine*, vol. 20, no. 1, p. 119, 1999.

[7] N. Noy, D. McGuinness *et al.*, "Ontology development 101: A guide to creating your first ontology," 2001.

[8] B. Chandrasekaran, J. Josephson, and V. Benjamins, "What are ontologies, and why do we need them?" *Intelligent Systems and Their Applications, IEEE*, vol. 14, no. 1, pp. 20–26, 2002.

[9] A. Maedche and S. Staab, "Ontology learning for the semantic web," *Intelligent Systems, IEEE*, vol. 16, no. 2, pp. 72–79, 2005.

[10] G. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[11] C. Hoege, B. Pfander, G. Moldovan, G. Pyrowolakis, and S. Jentsch, "RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO," *Nature*, vol. 419, no. 6903, pp. 135–141, 2002.

[12] C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira, "An introduction to the syntax and content of Cyc," in *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. Citeseer, 2006, pp. 44–49.

[13] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 697–706.

[14] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," *The Semantic Web*, pp. 722–735, 2007.

[15] J. Voss, "Collaborative thesaurus tagging the Wikipedia way," *Arxiv preprint cs/0604036*, 2006.

[16] M. Buffa and F. Gandon, "SweetWiki: semantic web enabled technologies in Wiki," in *Proceedings of the 2006 international symposium on Wikis*. ACM, 2006, pp. 69–78.

[17] M. Buffa, F. Gandon, G. Ereteo, P. Sander, and C. Faron, "Sweetwiki: A semantic wiki," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 1, pp. 84–97, 2008.

[18] M. Strube and S. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, no. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 1419.

[19] R. Mihalcea, "Using wikipedia for automatic word sense disambiguation," in *Proceedings of NAACL HLT*, vol. 2007, 2007.

[20] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer, "Semantic wikipedia," in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 585–594.

[21] G. Klyne, J. Carroll, and B. McBride, "Resource description framework (RDF): Concepts and abstract syntax," *Changes*, 2004.

[22] D. McGuinness, F. Van Harmelen *et al.*, "OWL web ontology language overview," *W3C recommendation*, vol. 10, pp. 2004–03, 2004.

[23] D. Brickley, R. Guha, and B. McBride, "RDF vocabulary description language 1.0: RDF schema," *W3C recommendation*, vol. 10, pp. 27–08, 2004.

[24] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: A large ontology from wikipedia and wordnet," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 3, pp. 203–217, 2008.

[25] F. Wu and D. Weld, "Autonomously semantifying wikipedia," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 41–50.

[26] D. Bonino, F. Corno, L. Farinetti, and A. Bosca, "Ontology driven semantic search," *WSEAS Transaction on Information Science and Application*, vol. 1, no. 6, pp. 1597–1605, 2004.

[27] (2010, November) London. Semantic MediaWiki. [Online]. Available: http://semanticweb.org/wiki/London

[28] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: Principles and state of the art," in *World Wide Web Conference*, 2008.

Appendix A: Reflection

In the aspect of web technology, the Semantic Web has become a significant topic. The first step to implement the Semantic Web is to build a large ontology based on real world knowledge system. My topic is to investigate some effective methodology to construct or refine the ontology from existing source such as Wikipedia or WordNet. The methodology in this topic mainly includes manual collaboration and machine automatic generating. So the chosen papers mainly concern on this two areas. Manual collaboration is not common in constructing ontology, because of high cost of human workforce and limitation of quality controlling. Therefore, this survey primary introduces the automatic way. The three key publications are comparable initiative in respective area. By reading the related work, I find some other projects or papers related to the area of my chosen paper which could provide the comparison with each other. The reference papers are mainly related to ontology reconstruction and information extraction from Wikipedia and WordNet by searching the Google Scholar and IEEE Xplore. Through reviewing the papers related to ontology reconstruction and information extraction, I realise that Semantic Web is a cross subject of AI, social science and web technology rather than a discipline of web technology. What the current semantic web need is to link the existing well-defined ontology rather than to create some new ontologies. Moreover, I have learned to use comparison and critical thinking in my survey.